STATS 370: Bayesian Statistics
Stanford University, Winter 2016

Problem Set 7

Solution provided by Ahmed Bou-Rabee

Problem 1. Multivariate Gaussian

Solution 1. We motivate our choice for the conjugate prior by first writing the likelihood function.

$$f(x_1, \ldots, x_n) \propto |\Sigma|^{-n/2} \exp\left(\frac{-1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)$$

Then, using $\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) = n(\mu-\bar{x})^T\Sigma^{-1}(\mu-\bar{x}) + \sum_{i=1}^{n}(x_i-\bar{x})^T\Sigma^{-1}(x_i-\bar{x})$ and the trace trick, we can write this as

$$f(x_1, \ldots, x_n) \propto |\Sigma|^{-n/2} \exp\left(-n/2(\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu)\right) \exp\left(-1/2\text{tr}(\Sigma^{-1}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^T)\right)$$

This looks like the product of a multivariate normal distribution of μ given Σ and a Wishart distribution of Σ. Looking on wikipedia, we see that this is the multivariate normal-Inverse-Wishart distribution. Specifically,

$$(\mu, \Sigma) \sim NW(\mu_0, k_0, \lambda_0, v_0)$$

where $\mu_0, k_0, \lambda_0, v_0$ are hyper-parameters of the normal-Inverse-Wishart distribution.

We now derive the joint posterior distribution associated to this prior. Let $(\mu, \Sigma) \sim NW(\mu_0, \lambda, W, v)$. And let $x_1, \ldots, x_n$ be the observations from $N(\mu, \Sigma)$. Then, the posterior

$$\begin{aligned}
p(\Sigma, \mu | x_1, \ldots, x_n) &\propto p(x_1, \ldots, x_n | \Sigma, \mu) p(\Sigma, \mu) \\
&\propto |\Sigma|^{-n/2} \exp\left(-n/2(\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu)\right) \exp\left(-1/2\text{tr}(\Sigma^{-1}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^T)\right) \\
&|\Sigma|^{((-(v_0+k)/2+1)} \exp\left(-1/2tr(\lambda_0\Sigma^{-1}) - k_0/2(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)\right)
\end{aligned}$$

Which we see is the normal-Wishart distribution with parameters $\mu_0* = \frac{k_0\mu_0+n\bar{x}}{k_0+n}$, $k_0* = k_0 + n$, $\lambda_0* = \lambda_0 + \sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T + k_0 n/(k_0+n)(\mu_0-\bar{x}))\mu_0-\bar{x})^T$, and $v_0* = v_0 + n$. This is the same family of the prior, and is hence conjugate.

Now we compute the marginal distribution of μ and we find that it is the multivariate student distribution. We do this by first computing the marginal of the joint prior and then substituting in the posterior parameters. By integrating out Σ, and using the fact that if A is a k by k non singular matrix and v is a k by k dimensional column vector, then $|A + vv^T| = |A|(1 + v^T A^{-1} v)$.

$$\begin{aligned}
\pi(\mu) &= \int_\Sigma \pi(\mu, \sigma) \\
&\propto \int_\Sigma |\Sigma|^{-(v_0+k)/2+1} \exp\left(-1/2tr(\lambda_0\Sigma^{-1}) - k_0/2(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)\right) \\
&\propto [\lambda_0 + (k_0+n)(m-\mu)(m-\mu)^T|^{-v_0+n+1)/2} \\
&\propto [1 + (k_0+n)(\mu-\mu_0)^T\lambda_0^{-1}(\mu-\mu_0)]^{-1/2(v_0+n+1)}
\end{aligned}$$

We recognize this as the multivariate t distribution. Therefore, the posterior marginal distribution of μ is a multivariate t distribution with mean μ_0^*, $v_0^* + n - k + 1$ degrees of freedom, and scale matrix $\lambda_0^*/(k_0^*(v_0^* - k + 1))$.

Problem 2. Collinearity

Solution 2. Let's assume that $\beta \sim N(0, \lambda I)$ for some $\lambda > 0$. Let $\Lambda^{-1} = \lambda I$. We know that $(y - x_i\beta) = \epsilon_i \sim N(0, \sigma^2)$. Thus, using Baye's rule, the posterior for β is

$$\begin{aligned} p(\beta|y, X, \sigma^2) &\propto p(\beta)p(y, X|\beta, \sigma^2) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) - \frac{1}{2}\beta^T\Lambda\beta\right) \end{aligned}$$

Let $\hat{\beta} = (X^TX + \sigma^2\Lambda)^{-1}X^Ty$. Then, we can write

$$\frac{-1}{\sigma^2}(y - X\beta) - \beta^T\Lambda\beta$$

as

$$(\beta - \hat{\beta})^T\frac{X^TX + \sigma^2\Lambda}{\sigma^2}(\beta - \hat{\beta}) + y^Ty - \hat{\beta}^T(X^TX + \Lambda)\hat{\beta} + \hat{\beta}^T\Lambda\hat{\beta}$$

Disregarding the components without β, and expanding we get,

$$p(\beta|y, X\sigma^2) \propto \exp\left((\beta - (X^TX + \sigma^2\Lambda)^{-1}X^Ty)^T\frac{X^TX + \sigma^2\Lambda}{\sigma^2}(\beta - (X^TX + \sigma^2\Lambda)^{-1}X^Ty)\right)$$

We recognize this as $p(\beta|y, X, \sigma^2) \sim N((X^TX + \sigma^2\Lambda)^{-1}X^Ty, \sigma^2(X^TX + \sigma^2\lambda)^{-1})$. We complete the regression analysis by choosing the mean of the posterior as our estimate. That is, we have $\hat{\beta} = X^TX + \sigma^2\Lambda)^{-1}X^Ty$, which we recognize as the ridge regression solution. Note that while (X^TX) does not have an inverse because of collinearity, $X^TX + \lambda I$ always does when $\lambda > 0$. And, when $n > k$ the solution still exists when $\lambda > 0$.

Problem 3. How to use the posterior distribution for point estimates

Solution 3.

1. This function is convex and differentiable so we differentiate with respect to $\hat{\theta}$ and then take the point where the derivative is 0. Thus, we have $2\theta = 2\hat{\theta}$. We then take expectations to get $\hat{\theta} = \mathbb{E}\theta$, which is the posterior mean.

2. Let $F(\theta)$ denote the cumulative posterior distribution function of θ. Then again to calculate the minimum we have to differentiate with respect to $\hat{\theta}$, but we can rewrite the minimizer first, (writing out the expected value in integral form)

$$\begin{aligned} &\frac{d}{d\hat{\theta}}\int_\Omega |\theta - \hat{\theta}|p(\theta)d\theta \\ &= \frac{d}{d\hat{\theta}}(\int_{\Omega(\theta>\hat{\theta})}(\theta - \hat{\theta})p(\theta)d\theta + \int_{\Omega(\theta<\hat{\theta})}(\hat{\theta} - \theta)p(\theta)d\theta) \\ &= \frac{d}{d\hat{\theta}}(\int_{\Omega(\theta>\hat{\theta})}\theta p(\theta)d\theta - \int_{\Omega(\theta<\hat{\theta})}(\theta)p(\theta)d\theta + \hat{\theta}F(\hat{\theta}) - \hat{\theta}(1 - F(\hat{\theta}))) \\ &= F(\hat{\theta}) - (1 - F(\hat{\theta})) \end{aligned}$$

Setting this equal to 0 and solving we get $F(\hat{\theta}) = \frac{1}{2}$, or $\hat{\theta}$ is the posterior median.

3. By definition, $\mathbb{E}L_0(\theta, \hat{\theta}) = P(|\theta - \hat{\theta}| > \epsilon) = 1 - P(|\theta - \hat{\theta}| < \epsilon)$. And then, taking $\epsilon \to 0$, we see this is $1 - P(\theta = \hat{\theta})$ and this value is minimized when $P(\theta = \hat{\theta})$ is as large as possible and this occurs when $\hat{\theta}$ is chosen to be the maximum of the posterior distribution, or the mode of $p(\theta)$.